PNAS
www.pnas.org

Supplementary Information for

Topography of epithelial-mesenchymal plasticity

Francesc Font-Clos, Stefano Zapperi and Caterina A. M. La Porta

Caterina A. M. La Porta.
E-mail: caterina.laporta@unimi.it

This PDF file includes:

Supplementary text
Figs. S1 to S10
Table S1
Captions for Databases S1 to S2
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Databases S1 to S2

Supporting Information Text

EMT-MET network. The EMT-MET network is build up from the EMT network in (1). We add the LIF to KLF4 pathway that goes through JAK-STAT signaling, see Fig. S1, as this allows for miR-200/SNAI1-induced MET under the original Boolean framework in (1).

Boolean model states. Each node in the model is considered to be either expressed or not expressed, without the possibility for intermediate states. The state of the system is encoded as a Boolean vector $\vec{s} = (s_1, s_2, \ldots, s_{72})$, where

$$s_i = \begin{cases} 1 & \text{if node } i \text{ is expressed} \\ -1 & \text{if node } i \text{ is not expressed} \end{cases} \quad [1]$$

Sets of promoters and inhibitors. Nodes that appear in the original Boolean rule preceded with the "not" operator are considered inhibitors, while the rest are considered promoters. This entails a set of promoters and a set of inhibitors for each node. As an example, the original Boolean rule for ZEB1 is `'ZEB1* = (HIF1 or SNAI1 or Goosecoid) and not miR200'`. Thus, the promoters of ZEB1 are HIF1, SNAI1 and Goosecoid, while its only inhibitor is miR200. The complete list of promoters and inhibitors of each node is provided as an excel file (`SI-Table-2.xlsx`).

Interaction matrix. The sets of promoters and inhibitors are coded into an interaction matrix J_{ij}, where

$$J_{ij} = \begin{cases} 1 & \text{if node } j \text{ promotes node } i \\ -1 & \text{if node } j \text{ inhibits node } i \\ 0 & \text{otherwise} \end{cases} \quad [2]$$

Parsing the original Boolean rules in (1) as explained above yields a total of 142 interactions. Promoting interactions turn out to be about three times more common than inhibiting interactions ($n_{\text{prom}} = 107, n_{\text{inh}} = 35$).

Boolean dynamics. Nodes are updated following a majority rule:

$$s_i(t+1) = \text{sign}\left(\sum_j J_{ij} s_j(t)\right). \quad [3]$$

In the case of $\sum_j J_{ij} s_j = 0$, the node is not updated, keeping its present state (2). Readers with some background in statistical physics will recognize Eq. (3) as the zero temperature Glauber dynamics of an disordered Ising model commonly used to model ferromagnetic hysteresis in ferromagnets (3) and spin glasses (4) It is easy to get an intuitive feel of how nodes are updated: first, the sum over j in Eq. (3) is only effective when $J_{ij} \neq 0$, so that the state of a node depends only on the current state of its promoters and inhibitors, as it should. Indeed, we can rewrite Eq. (3) as

$$s_i = \text{sign}\left(\sum_{j\in\mathcal{P}} s_j - \sum_{j\in\mathcal{I}} s_j\right) \quad [4]$$

where $\mathcal{P}$ denotes the set of promoters of node i and $\mathcal{I}$ its set of inhibitors. Using the substitution $s_i = 2z_i - 1$, we get

$$z_i = \Theta\left(\sum_{j\in\mathcal{P}} z_j - \sum_{j\in\mathcal{I}} z_j - f_i\right) \quad [5]$$

with $f_i = 1/2(|\mathcal{P}| - |\mathcal{I}|)$. Thus a node becomes active (inactive) if the difference between the number of active promoters and active inhibitors is more (less) than a threshhold f_i that depends only on the topology of the network and not on its state. The variable f_i corrects for the bias that would otherwise appear due to unequal number of promoters/inhibitors of each node. Formally, one can show that if z_j are unbiased Bernoulli random variables, so it is z_i, for any interaction matrix J_{ij}. This property is the equivalent of the *half-functional rule* in (5), where Huang et. al. study the robustness of continuous-state regulatory circuits by randomizing the parameters of a system of differential equations. Throughout the present work, nodes in the network are updated asynchronously as in (1), but since we are only interested in steady-states we do not account for different speeds of regulatory processes as in (1).

Network reconstruction errors and the half-functional rule. The reconstruction of gene regulatory networks is known to be a delicate job, and even when extreme care is taken during reconstruction, as in (1), there is always the possibility that real but unobserved interactions take place, or that some of the inferred interactions are in reality technical or methodological artifacts. In our framework, network reconstruction errors can be modeled by changing the value of f_i, as follows:

$$f_i = 1/2(|\mathcal{P}| - |\mathcal{I}|) + h_i \quad [6]$$

The external local field h_i biases the node towards ON/OFF states, while the case $h_i \equiv 0$ corresponds to the unbiased scenario. Beyond the fact that network reconstruction errors are inaccessible to us by definition – so that choosing h_i in a sensible way is

highly challenging– the *half-functional rule* (5) predicts that nodes with large biases are effectively frozen in continuum-state models. Genes whose expression does not change should not be part of the network reconstruction. Thus, it is reasonable to set external local fields to zero, at least for the continuum case. To test if this reasoning extends to our discrete setting, we implement two versions of the model with random local fields, $h_i \in \{-\Delta, 0, \Delta\}$ with equal probabilities and $\Delta = \epsilon = 10^{-3}$ or $\Delta = 1$. Notice that the two cases are conceptually different, as $\Delta = \epsilon \ll 1$ effectively can only "break ties", while $\Delta = 1$ is a stronger bias, but allows for ties as in the original model. Figure S8 shows clustering of steady states for the random-field model, as well as the original model for comparison. We also measure the ratio of frozen nodes, defined as those that are always ON or always OFF in the sampled steady states. We find that adding random-fields leads to a large and variable proportion of nodes being frozen. These fluctuations depend on the disorder realization (the values of the random-fields h_i), and render the modified models less attractive from the biological point of view: large portions of the network reconstruction are frozen, contradicting available experimental data.

Calculation of the pseudo-Hamiltonian. The pseudo-Hamiltonian H associated to a steady state $\vec{s}$ is defined as

$$H = -\sum_{i,j} J_{ij}\, s_i s_j \qquad [7]$$

In statistical physics, this quantity translates into the Hamiltonian of an asymmetric random bond Ising model. Since the majority of the coupling constants J_{ij} is positive, the model is effectively a random ferromagnet and not a spin glass, where $\langle J_{ij} \rangle = 0$. The variable H measures how much a given steady state respects the constraints imposed by J_{ij}, and assigns a value of energy accordingly: states that satisfy more interactions get lower H values, while states that leave more frustrated interactions have higher H values.

Stochastic decrease of H. The change of H under the update rule Eq. (3) is guaranteed to be negative for symmetric matrices $J_{ij} = J_{ji}$, but the asymmetric case is more involved, see (6, 7). Here, we first provide some general derivations and then empirically measure the probability by which H decreases over time in our network. Assume that node k changed state at time-step t after application of Eq. (3), i.e.

$$s_k(t+1) = \operatorname{sgn} \sum_j J_{kj} s_j(t) \qquad [8]$$

$$s_k(t) = -s_k(t+1) \qquad [9]$$

Introducing the per-node input and output energies $\rho_k^{\text{in}} = \sum_j J_{kj} s_j(t)$, $\rho_k^{\text{out}} = \sum_i J_{ik} s_i(t)$, it is easy to see that $\Delta H = H_{t+1} - H_t = -2\ \operatorname{sgn}(\rho_k^{\text{in}})\ [\rho_k^{\text{in}} + \rho_k^{\text{out}}]$. Denoting $c_k = 2|\rho_k^{\text{in}}| > 0$, we can rearrange the former expression into

$$\Delta H = -c_k \left(1 + \frac{\rho_k^{\text{in}}}{\rho_k^{\text{out}}}\right), \qquad [10]$$

which sets a simple condition for the sign of ΔH. At this point it is instructive to inspect two special cases: if $J_{ij} = J_{ji}$, then $\rho_k^{\text{in}} = \rho_k^{\text{out}}$ and $\text{Prob}(\Delta H \leq 0) = 1$; while if $J_{ij} = -J_{ji}$, $\rho_k^{\text{in}} = -\rho_k^{\text{out}}$ and $\text{Prob}(\Delta H = 0) = 1$. When neither of these cases is satisfied, the sign of ΔH cannot be determined a priori and must be empirically measured. For our model, we find $\text{Prob}(\Delta H \leq 0) \simeq 0.87$, with $\langle \Delta H \rangle = -1.86$. Figure S9b shows the evolution of H as a function of t, from the initial random configuration until a steady state is reached, for a set of 20 randomly chosen realizations of the model. As expected, the value of H tends to decrease over time, drifting the system towards low-H states. In summary, our interaction network is such that, overall, H decreases under the iterative application of Eq. (3), even if there is a small but finite probability of $\Delta H > 0$ at the level of a single time-step.

Calculation of state overlaps $q_{\alpha\beta}$. The overlap between two steady states $\vec{s}_\alpha, \vec{s}_\beta$ is defined as

$$q_{\alpha\beta} = \frac{1}{N} \sum_i (\vec{s}_\alpha)_i\ (\vec{s}_\beta)_i \qquad [11]$$

where N denotes the number of nodes of the network, and $(\vec{s}_\alpha)_i$ denotes the i-th component of the steady state labeled α. Notice that $q_{\alpha\beta} = 1$ if $\vec{s}_\alpha = \vec{s}_\beta$, and $q_{\alpha\beta} = -1$ if $\vec{s}_\alpha = -\vec{s}_\beta$. Thus the overlap is a similarity measure over Boolean states, from -1 (exactly opposite state) to $+1$ (exactly equal states).

Sampling of steady states and calculation of abundances. A network state $\vec{s}$ is considered to be a steady state if all of its nodes remain unchanged when they are updated via Eq. (3). To sample different steady states, we start from different random initial conditions and update the state of nodes nodes as explained above until a steady state $\vec{s}_*$ is reached. In the present asynchronous model, we never find limit cycles, which are prevalent in random Boolean networks under periodic update (8, 9).

$$\vec{p}_0 \xrightarrow{s_i = \operatorname{sign}\left(\sum_j J_{ij} s_j\right)} \vec{p}_1 \xrightarrow{s_k = \operatorname{sign}\left(\sum_j J_{kj} s_j\right)} \vec{p}_2 \to \cdots \to \vec{s}_* \qquad [12]$$

Starting with a total of 10^7 different random initial conditions, we end up with 1 198 287 different steady states. We define the relative abundance $a(\vec{s})$ of a steady state $\vec{s}$ as the probability of finding $\vec{s}$ starting from random initial conditions. Numerically, we estimate $a(\vec{s})$ simply as

$$a(\vec{s}) = \frac{\#\{\text{Simulations that lead to steady state } \vec{s}\}}{\#\{\text{Simulations}\}}. \quad [13]$$

To each steady state $\vec{s}$ corresponds a value of a. The quantity $P(a)$ in Fig. 1 is then the distribution of relative abundances $a(\vec{s})$ over a set of steady states.

Simulation of standard and transient KD/OE conditions. Knock-down (KD) (over-expression (OE)) conditions are modeled by fixing the state of a node to -1 ($+1$). Under standard KD/OE conditions, the state of such a node cannot change during the update process. We also consider transient OE/KD conditions, where the node being initially switched is allowed to eventually relax back to its initial state. Steady-states whose energy is plotted in Figure 2(a,b) are computed taking random initial conditions with SNAI1 locked into -1 (KD) or $+1$ (OE). To simulate SNAI1-induced EMT in Figure 2(c) we perturb WT steady states by first locking SNAI1 into $+1$ (OE), and then running the model until a new steady state is found.

Computation of marker state probabilities. In Figures 1b,c; 3b,c; 4a-d; S4; S5; S6 and S7; model estimates of the probabilities of specific markers to be expressed (state=ON) are reported using different color schemes on the PCA plane of steady states . These probabilities are computed as the ratio of steady states where the marker is ON in a local neighborhood of a given (x, y)-point in PCA space, obtaining thus a number $0 \leq r_{(x,y)} \leq 1$.

Computation of transition probabilities. EMT/MET transition probabilities are computed using 10^4 steady states. We compute separately EMT/MET transition probabilities due to OE/KD. In this way, to compute the EMT probability via SNAI1 OE, we must consider the subset of steady states that express E-cadherin but do not express SNAI1. For each of these steady states, we set and lock SNAI1 to $+1$, simulating OE conditions, and run the model dynamics until a new steady state is found. The ratio of simulations where the final steady state does not express E-cadherin is the estimated EMT probability via SNAI1 over-expression.

Simulation of avalanches. We consider avalanches triggered by both permanent and transient OE/KD. The size of the avalanche is the total number of nodes that have changed due to the perturbation. The distribution of avalanches is computed using 10^5 steady states.

Finite-size effects in the GTEx dataset. Computing state abundances $a(\vec{s})$ using the full 72 nodes in the GTEx dataset leads to $a(\vec{s}) = 1$ for almost all states $\vec{s}$. This is because the number of possible states ($2^{72} \simeq 10^{21}$) is extremely large, compared with the available number of samples (11688), leading to insufficient statistics. To circumvent this issue, we restrict our analysis to a subset of $N < 72$ nodes (sorting them by their binarized significance). We systematically study how the number of unique states depends on the number of nodes N when the number of samples is fixed to $S = 11688$, comparing the model, the GTEx dataset and an additional random dataset where all nodes are independent from each other. Figure S3a shows that the model and the GTEx dataset behave similarly and differ from the random dataset in a range of intermediate values of N. For very high or very low values of N, however, finite-size effects dominate, and the model and the GTEx dataset cannot be distinguished from a random dataset. To choose a non-arbitrary intermediate value of N, we define N_{opt} as the value of N that minimizes the ratio of unique steady states between the model and the random dataset. Figure S3c shows that N_{opt} increases with increasing sample size S. Following this analysis, state abundances in Figure 3d are estimated using $N_{\text{opt}} = 14$.

Alternative fits for distributions of abundances and avalanches. We consider a power law with cutoff, with probability density function

$$f_1(x, \alpha, \lambda) \propto \frac{e^{-(x/\lambda)^2}}{x^\alpha}$$

and a log-normal distribution, with probability density function

$$f_2(x, s, \mu) \propto \frac{1}{x} \exp\left(-\frac{1}{2}\left(\frac{\log(x) - \mu}{s}\right)^2\right)$$

as alternative functional forms that can account for finite-size effects in the distributions of abundances and avalanches. Using mean-squared-error (MSE) minimization on the logarithm of the density functions over logarithmically-spaced values of x, we find the MSE parameters indicated in Table S1. Figure S10 shows the survival function of the resulting fits for the distribution of abundances of the model, the distribution of abundances of the GTEx dataset, and the avalanche size distributions of the model.

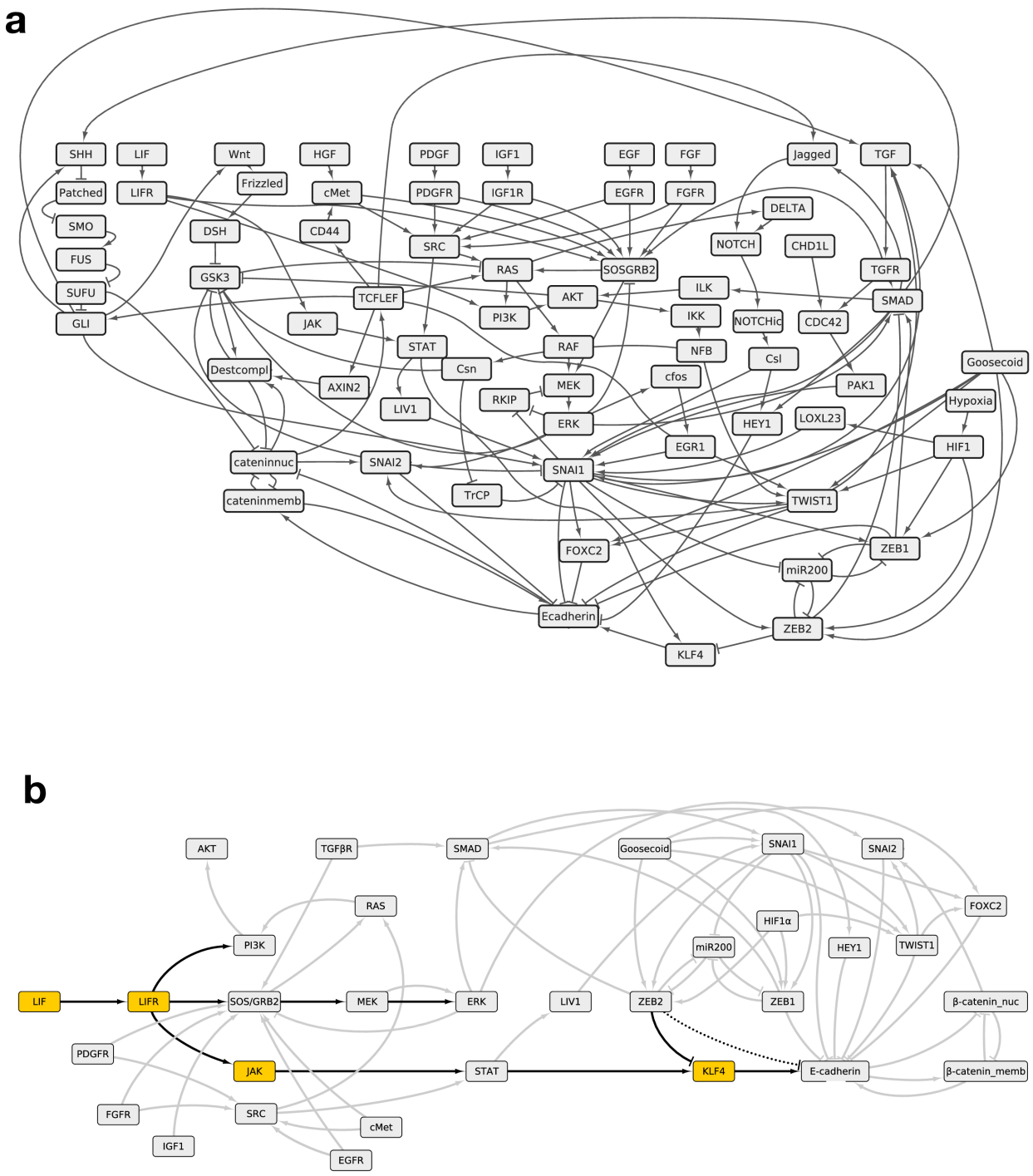

Fig. S1. The EMT-MET interaction network. (a) The full EMT-MET network. Promoting interactions are drawn with an arrow, while inhibiting interactions are drawn with a cap. (b) The original ZEB2-E-cadherin inhibiting interaction(1) (dotted black edge) is replaced by a more detailed scheme: ZEB2 inhibits KLF4, and KLF4 induces the expression of E-cadherin. The full JAK-STAT signaling pathway is also added, and connected to elements present in the original network. New nodes are drawn in yellow; edges corresponding to new interactions in black, and nodes and edges already present in the original model are drawn in gray. Only nodes at topological distance 2 or less from a new (yellow) node are shown.

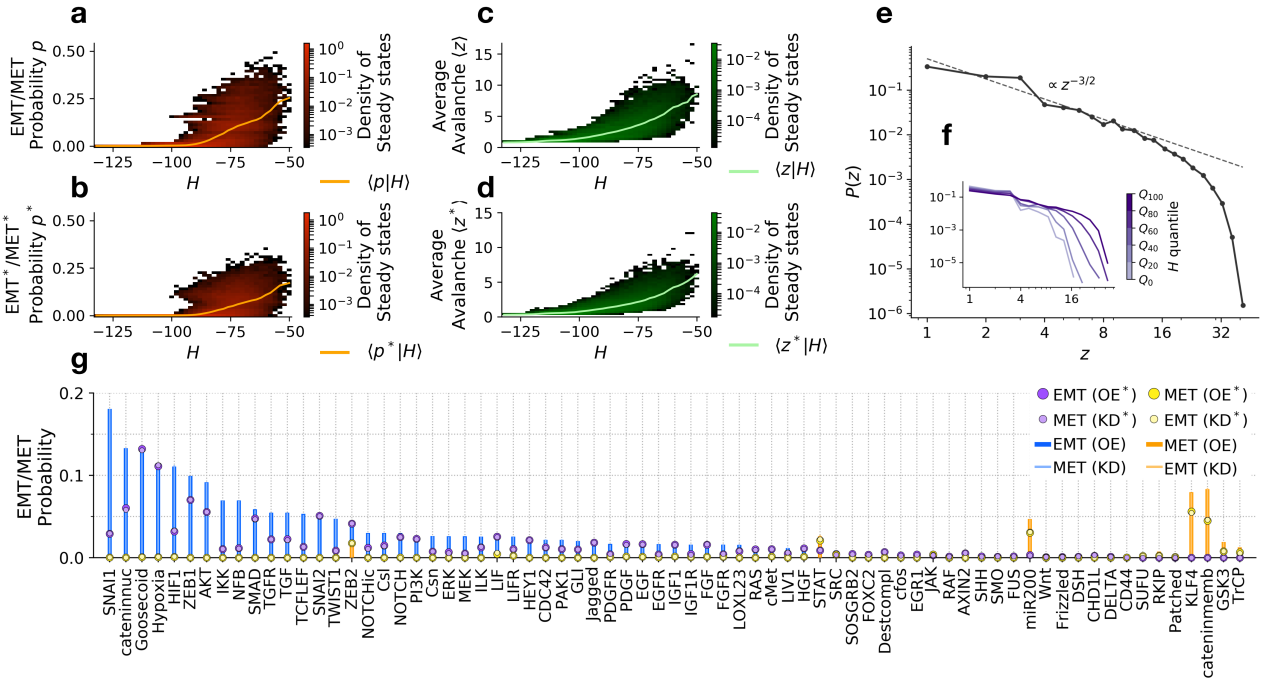

Fig. S2. $\mathbf{EMT/MET}$ probability and avalanches under stable and transient $\mathbf{OE/KD}$ conditions. (a, b) EMT/MET probability as a function of H, under (a) stable OE/KD or (b) transient OE/KD. (c, d): Average avalanche sizes $\langle z \rangle$ as a function of H, under (c) stable or (d) transient OE/KD conditions. (e) Avalanche size distribution under single-node stable OR (solid black line). The dashed black like of slope $\tau = 3/2$, expected for mean-field avalanches, is shown as a guide to the eye. The inset (f) shows that high-H states tend to give rise to larger avalanches. (g) EMT/MET probabilities under stable and transient KD/OE conditions. As expected, transient perturbations yield lower EMT/MET probabilities than stable perturbations.

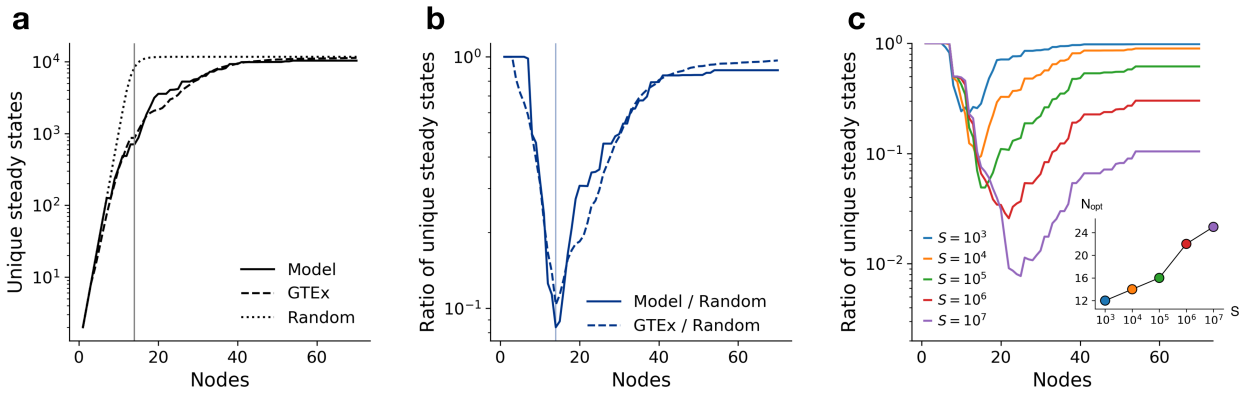

Fig. S3. $\mathbf{Finite\text{-}size\ effects\ in\ GTEx\ data}$ (a) Number of unique steady states as a function of number of nodes included in the analysis at fixed $S = 11688$ total steady states, matching the number of samples in the GTEx dataset. The model (solid black line) and the GTEx data (dashed black line) display very similar curves, and differ from what a random null model would yield (dotted black line). For very small or very large number of nodes, however, all three curves coincide due to finite-size effects. (b) Same as (a), showing the ratio of unique steady states with respect to the random null model, both for our model (solid blue line) and for the GTEx data (dashed blue line). A ratio of 1 indicates that finite-size effects dominate the statistics, making impossible to distinguish real data from a random null model. The optimal number of nodes $N_{\rm opt}$ is chosen such that the ratio of unique steady states is minimized. (c) Ratio of unique steady states of the model with respect to a random null model, for increasing sample sizes $S = 10^3, 10^4, \ldots, 10^7$. Inset: the value of $N_{\rm opt}$ increases linearly with the logarithm of S, showing that for a large enough sample size, eventually all nodes could be used.

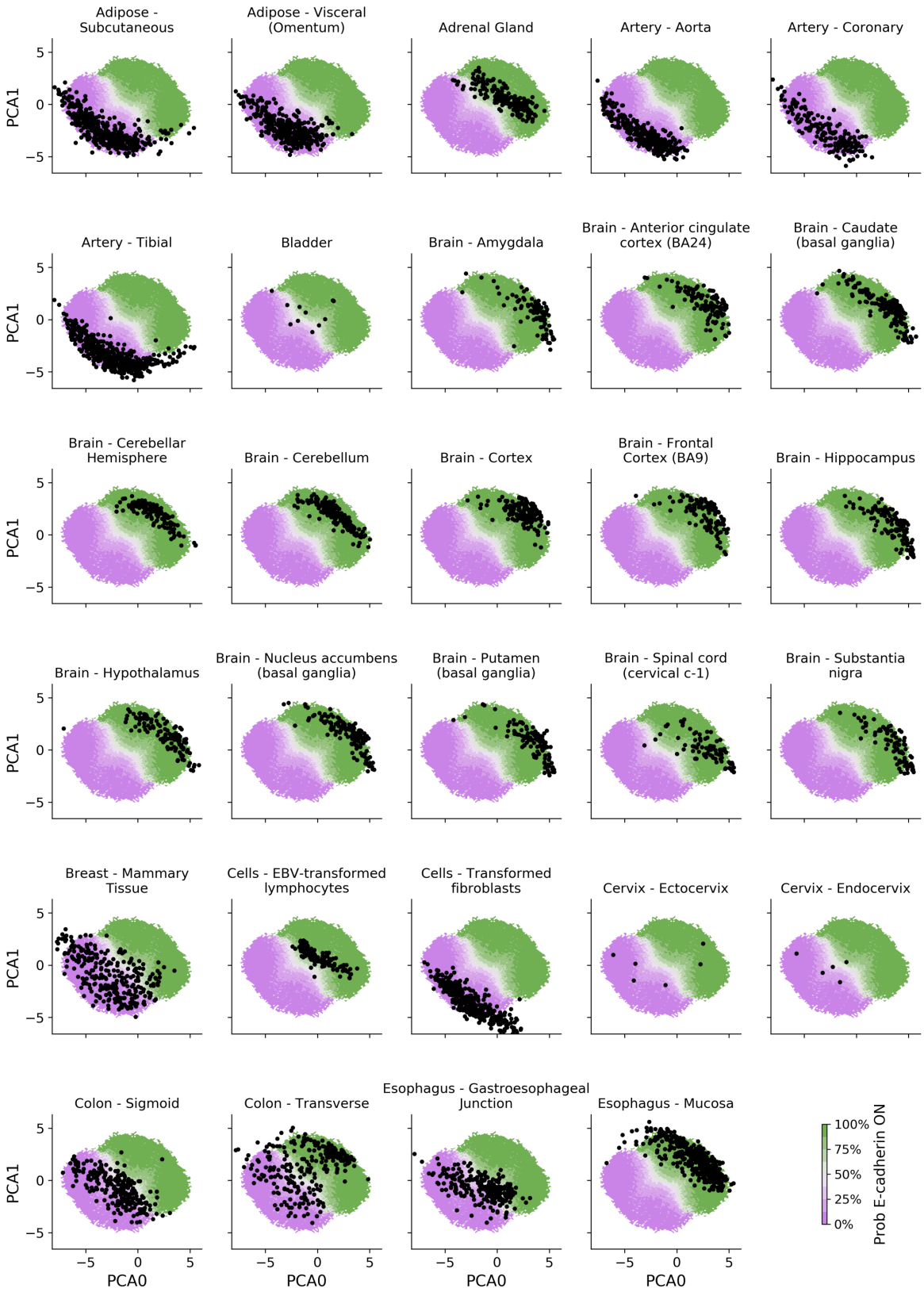

Fig. S4. GTEx tissues show different location in PCA space (1 of 2) Black scatter dots are GTEx samples. Each panel includes samples from a different tissue. The background coloring corresponds to the probability of E-cadherin being expressed in the model.

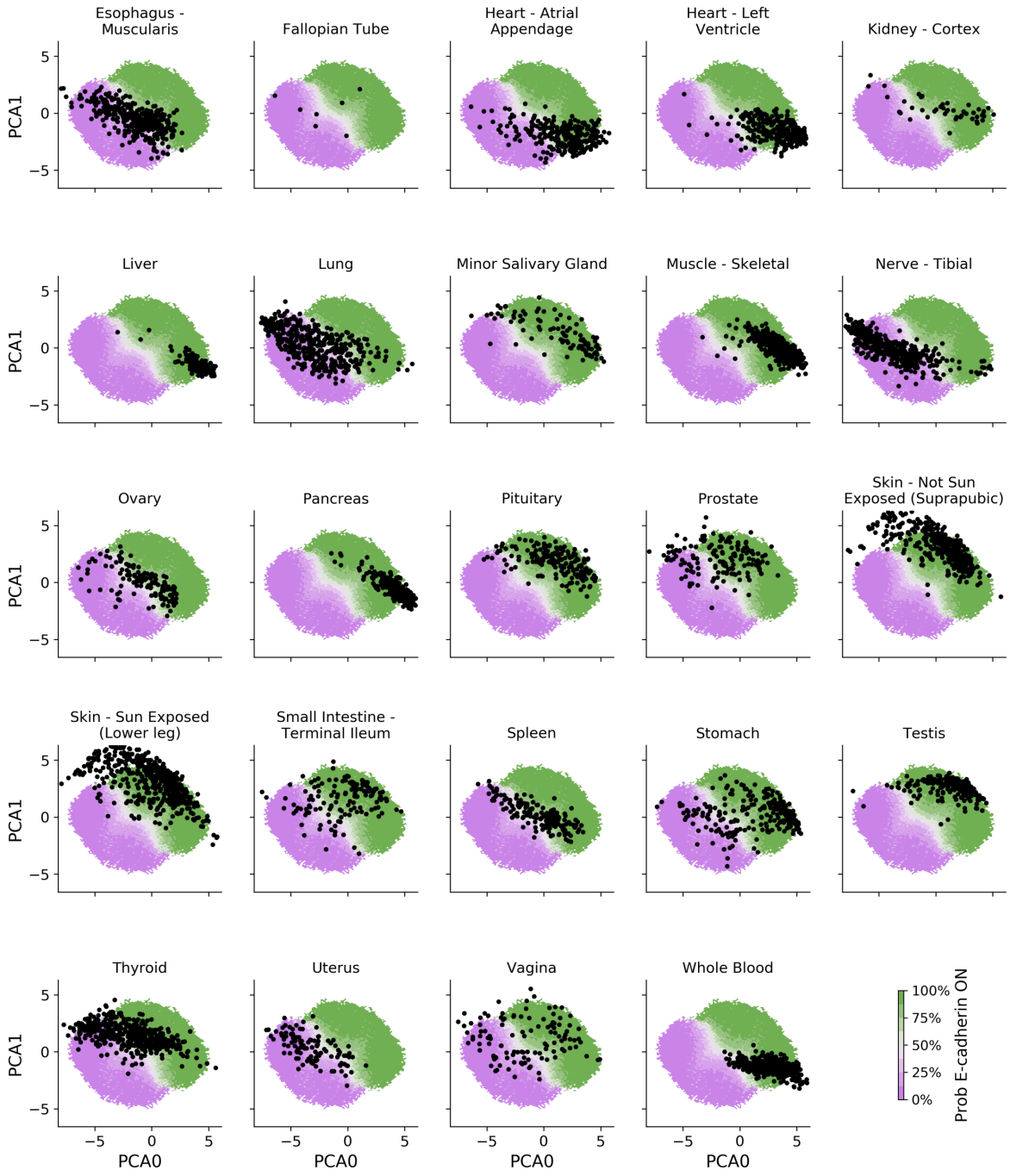

Fig. S5. **GTEx tissues show different location in PCA space (2 of 2)** Black scatter dots are GTEx sample. Each panel includes samples from a different tissue. The background coloring corresponds to the probability of E-cadherin being expressed in the model.

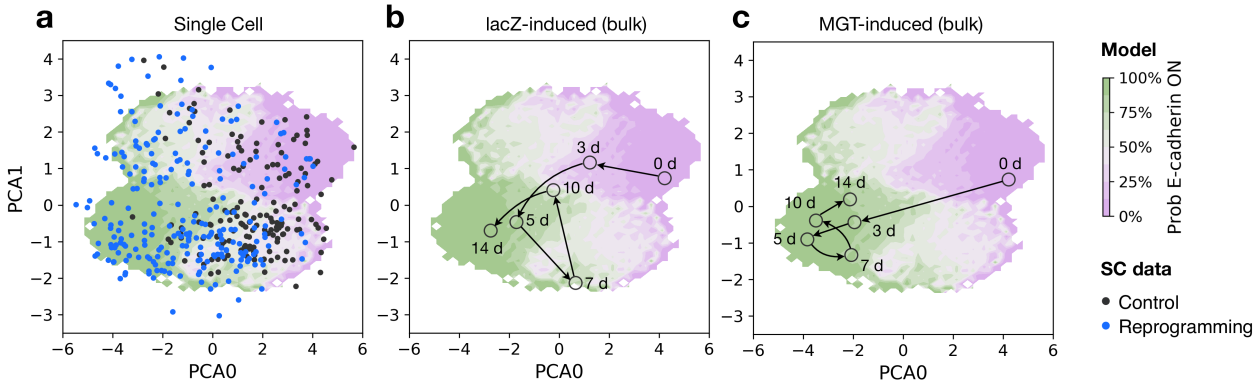

Fig. S6. Fibroblast-to-cardiomyocyte bulk and single-cell data. (a) Single-cell data for reprogramming cells (blue dots) and control cells (black dots) projected on top of the model PCA space. The background coloring shows the probability of CDH1 being expressed in the model. Reprogramming cells tend to lie on regions of high CDH1 probability (green), while control cells stay on low CDH1 probability regions (violet), signaling a possible M to E transition. (b, c): Time-course bulk data under lacZ (b) or MGT (c) treatment. In both cases, a transition from M to E states is clear, with MGT-induced cells displaying a faster transition than lacZ-induced ones. Data are obtained from GSE98570 (bulk data) and GSE98567 (single-cell data) (10).

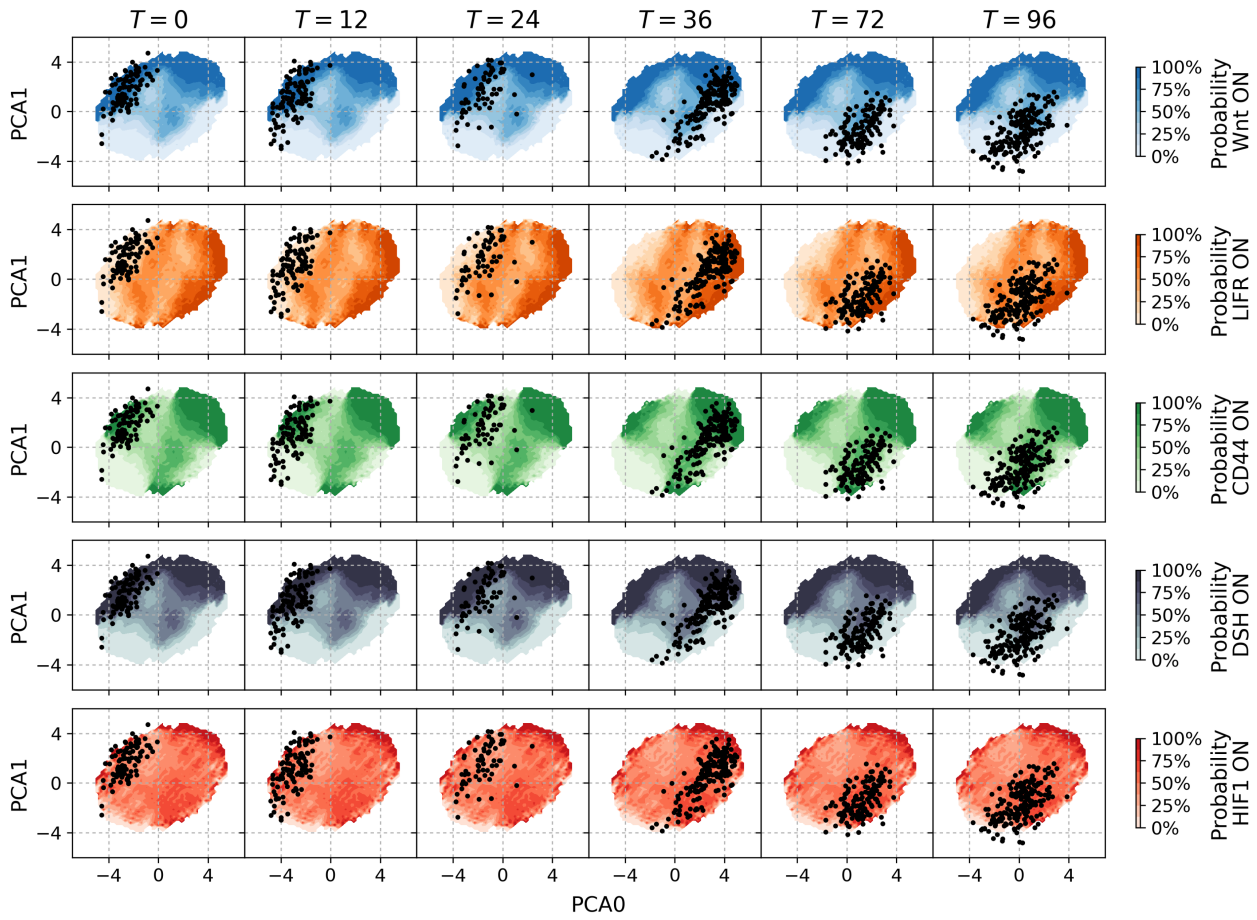

Fig. S7. Topographic maps according to different markers. Experimental data from single-cell embryonic-to-endoderm differentiation (GSE75748 (11)) move across the map as cells undergo EMT. The background color indicates the ratio of steady states that express a set of different markers: WNT, LIFR, CD44, DSH, HIF1. See also Fig. 4c.

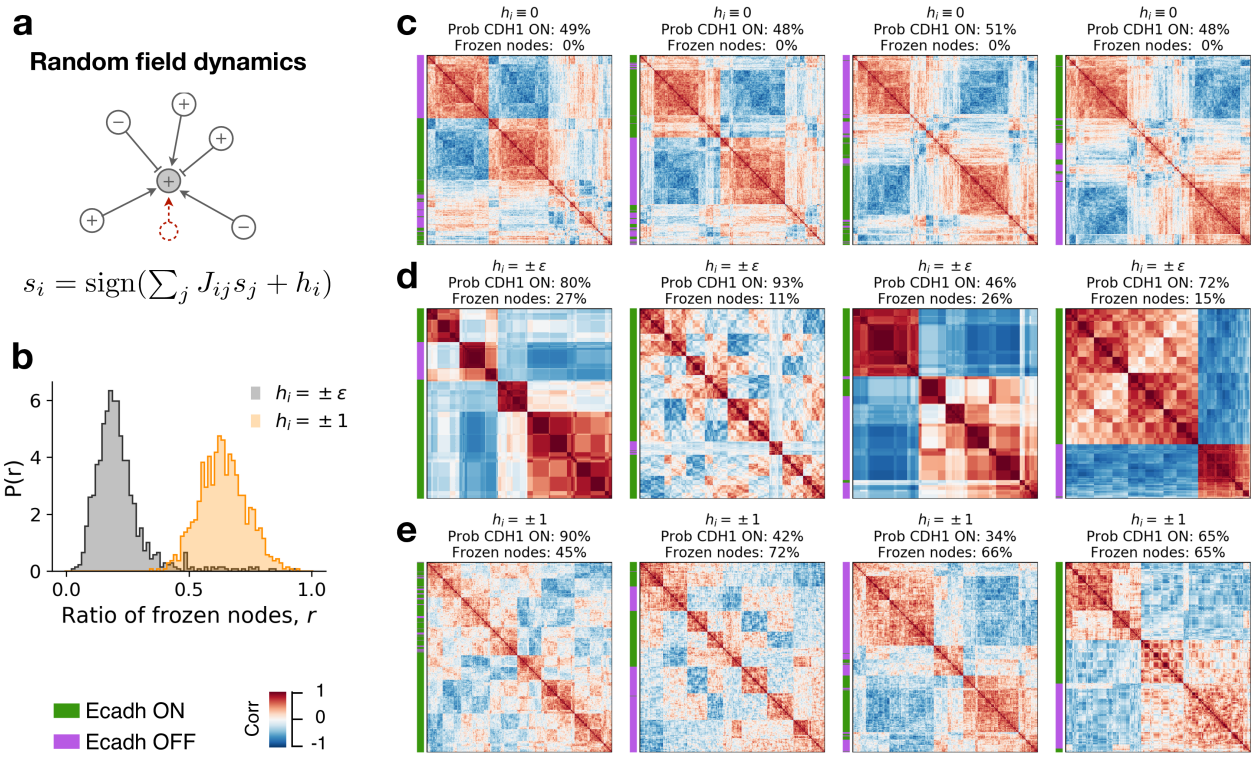

Fig. S8. **Random fields model** (a) Illustration of the random fields model, where nodes can be biased by an amount h_i due to e.g. network reconstruction errors. (b) The ratio of frozen nodes r for the two versions of the random fields model, $h_i = \pm\epsilon$ (gray) and $h_i = \pm 1$ (orange). The original model has by construction $P(r) = \delta(r)$. The panel shows that adding random fields freezes a large proportion of nodes. (c, d, e) Clustering of steady states, computed using 500 steady states of the model. The heatmap shows correlation between steady states. Colors on the left of each heatmap mark the expression of E-cadherin (green) or lack of expression (violet). We show the original model $h_i \equiv 0$ in panel (c) for comparison, and two versions of the random fields model: $h_i = \pm\epsilon$ (d) and $h_i = \pm 1$ (e). The ratio of frozen nodes and the probability of E-cadherin to be expressed are also shown. In panels (d, e), each heatmap corresponds to one realization of the disorder. The panels show that the hierarchical organization of steady states depends on the realization of the disorder.

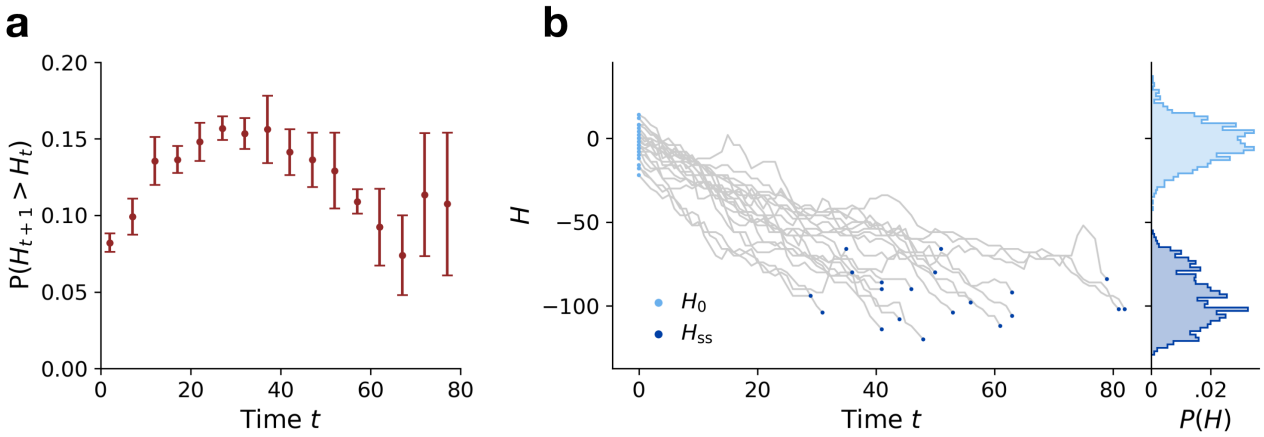

Fig. S9. **Probabilistic decrease of H** (a) The probability of $H_{t+1} > H_t$ as a function of t, for $N = 10^3$ realizations of the model. Data is binned in intervals of length 10. Error bars correspond to one standard deviation. (b) Histograms of H_0 (light blue) and $H_{\rm ss}$ (dark blue). $H_{\rm ss}$ denotes the value of H when a steady state is reached. Gray lines display the evolution of H_t for 20 randomly chosen realizations. A light (dark) blue dot marks the initial (final) value of H for each realization.

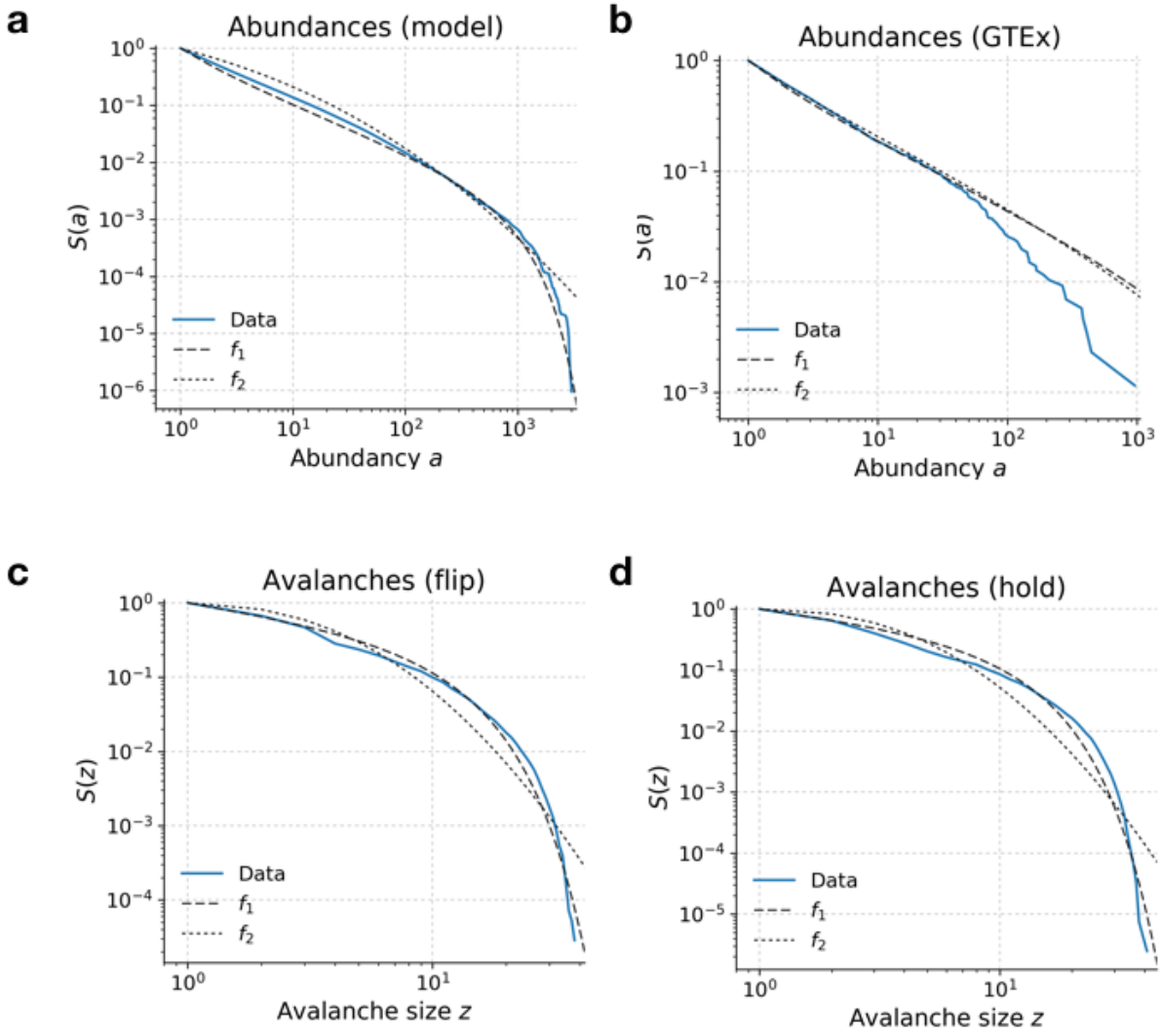

Fig. S10. Survival functions of abundances and avalanches The survival function $S(x) = \mathrm{Prob}[X \geq x]$ for the distribution of abundances in the model (a) and in the GTEx data (b), as well as for the distribution of avalanches under transient perturbations (c) or KD/OE (d). The blue solid curves display the empirical survival function, while the black dashed line corresponds to best MSE fit with a truncated power law f_1 and the black dotted one to the best MSE fit of a lognormal distribution (f_2).

	α	λ	s	μ
Abundances (model)	1.82	$10^{3.19}$	2.02	-0.06
Abundances (GTEx)	1.60	$10^{6.6}$	5.87	1.10
Avalanches (flip)	1.07	$10^{1.2}$	0.76	1.10
Avalanches (hold)	1.05	$10^{1.17}$	0.71	-17

Table S1. Alternative functional forms for the distributions of abundances and avalanches. The table shows the MSE parameters for $f_1(x, \alpha, \lambda)$ and $f_2(x, s, \mu)$

Additional data table S1 (SI-Table-1.xlsx)
Conversion tables between nodes in the networks and gene in transcriptomes.

Additional data table S2 (SI-Table-1.xlsx)
List of promoters and inhibitor for each node in the network.

References

1. Steinway SN, et al. (2014) Network modeling of tgf-beta signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint sonic hedgehog and wnt pathway activation. *Cancer Research* 74(21):5963–5977.
2. Li F, Long T, Lu Y, Ouyang Q, Tang C (2004) The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci. U. S. A.* 101(14):4781–4786.
3. Sethna JP, et al. (1993) Hysteresis and hierarchies: Dynamics of disorder-driven first-order phase transformations. *Phys. Rev. Lett.* 70(21):3347–3350.
4. Pázmándi F, Zaránd G, Zimányi GT (1999) Self-organized criticality in the hysteresis of the sherrington-kirkpatrick model. *Phys. Rev. Lett.* 83(5):1034–1037.
5. Huang B, et al. (2017) Interrogating the topological robustness of gene regulatory circuits by randomization. *PLOS Computational Biology* 13(3):1–21.
6. Derrida B (1987) Dynamical phase transition in nonsymmetric spin glasses. *J. Phys. A Math. Gen.* 20(11):L721.
7. Gutfreund H, Reger JD, Young AP (1988) The nature of attractors in an asymmetric spin glass with deterministic dynamics. *J. Phys. A Math. Gen.* 21(12):2775.
8. Gershenson C (2004) Introduction to random boolean networks in *Workshop and Tutorial Proceedings, Ninth International Conference on the Simulation and Synthesis of Living Systems*, ed. Bedau, M., P. Husbands, T. Hutton, S. Kumar, and H. Suzuki. pp. 160–176.
9. Harvey I, Bossomaier T (1997) Time out of joint: Attractors in asynchronous random boolean networks in *Proceedings of the Fourth European Conference on Artificial Life*. pp. 67–75.
10. Liu Z, et al. (2017) Single-cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte. *Nature* 551(7678):100–104.
11. Chu LF, et al. (2016) Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology* 17(1).